MOSQUITO CHRONOLOGICAL AGE DETERMINATION USING MID-INFRARED SPECTROSCOPY AND CHEMOMETRICS

A thesis presented to the faculty of the Graduate School of Western Carolina University in partial fulfillment of the requirements for the degree of Masters of Science in Chemistry.

By

Bradley Forrest Guilliams

Adviser: Dr. Scott Huffman
Associate Professor of Chemistry
Department of Chemistry & Physics

Committee Members: Dr. Brian Byrd, Environmental Health Sciences Program
Dr. Carmen Huffman, Department of Chemistry & Physics

April 2020

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ae.	*Aedes*
AMCA	American Mosquito Control Association
ANN	artificial neural networks
BEI	Biodefense and Emerging Infections Research Resources Repository
CA	California
Cx.	*Culex*
DRIFTS	diffused reflectance infrared fourier-transform spectroscopy
HDF	hierarchical data format
IMM	integrated mosquito management
IR	infrared
JDX	EDICT Index. J Database Exchange
MALDI-TOF MS	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MCT/A	mercury cadmium telluride
MR4	Malaria Research and Reference Reagent Resource Center
MSU	Michigan State University
NIAID	National Institute of Allergy and Infectious Diseases
NIH	National Institute of Health
NIR	near-infrared
PCA	principal components analysis
PCAR	principal components analysis regression
PLSR	partial least squares regression
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
SEPsv	standard error of prediction, separate validation
SPA	single-page application
SVM	support vector machines
WCU	Western Carolina University

ABSTRACT

MOSQUITO CHRONOLOGICAL AGE DETERMINATION USING MID-INFRARED SPECTROSCOPY AND CHEMOMETRICS

Bradley Guilliams, Masters of Science in Chemistry

Western Carolina University (April 2020)

Adviser: Dr. Scott Huffman

Determining a mosquito population's species composition and age is crucial for estimating the risk of pathogen transmission. At present, age-grading methods are chiefly physiologic and classify the mosquitoes in terms of parity (e.g., nulliparous or parous). Less commonly used chronologic methods (e.g., qPCR or near infrared spectroscopy [NIR]) have limited temporal resolution (NIR) or require consumable reagents and technological expertise with molecular methods. The current lack of robust methods to rapidly evaluate a population's chronologic age limits our ability to assess pathogen transmission risk in the context of vectorial capacity estimations (i.e., daily survivability). Our current research seeks to develop methods of mosquito age determination utilizing mid-infrared spectroscopy and advanced numerical analysis (chemometrics). Infrared (IR) spectroscopy is a type of vibrational spectroscopy that is both sensitive and information rich. Subtle changes in IR spectra correlate with changes in the biochemistry of mosquitoes as they age. It has been shown that mosquito species can be identified using mid infrared spectroscopy and chemometrics. Using mid-infrared spectroscopy and chemometrics, the chronologic age of *Aedes triseriatus* mosquitoes were predicted using PLSR and ANN models. *Aedes triseriatus* were successfully reared into groups of different ages

with low uncertainty in the age. *Aedes triseriatus* spectra were used to create a training dataset and fit models for prediction using PLSR and ANN. PLSR and ANN models were used to predict the age of samples using a validation dataset with SEPsv of 4.3 and 3.3 days respectively. Mean spectra for each age group were used to try and discern a specific chemical underpinning for the performance of these models and to explain why mosquito age could be predicted using PLSR and ANN models. Peaks between 1200 – 1000 cm^{-1} typically associated with chitin were investigated and the second derivative of mean absorbance by age at 1032 cm^{-1} increased linearly with age.

CHAPTER ONE: BACKGROUND

1.1 Importance of Mosquito Surveillance

Mosquito-borne diseases are responsible for significant morbidity and mortality at both an international and a domestic scale.[1,2] Tremendous economic costs can be attributed to the burdens of malaria, dengue, and to the growing threat of arboviruses such as the West Nile, chikungunya, and Zika viruses.[1,3] Many of these diseases do not have an effective vaccine, so mosquito control efforts such as source reduction, larvicides, and adulticides are used for arboviral disease prevention and control. Integrated Mosquito Management (IMM) as described by the American Mosquito Control Association (AMCA) includes surveillance to first assess the threat and several control efforts to reduce the mosquito population.[2,4] Mosquito identification has been a successful approach to understanding the mosquito population in a given area.[5,6] Frequent surveillance is paramount in determining the human health risk of a mosquito population. Population surveillance is especially important for mosquito species such as certain *Culex* and *Aedes* species that are capable of transmitting Zika,[7,8] dengue,[9] West Nile,[10] La Crosse,[11] Chikungunya,[12] or other arboviruses. Furthermore, understanding the age of a mosquito population gives mosquito control districts a more thorough understanding of the threat to human health for these mosquito species of vectorial importance. A glossary of biologically relevant terms can be found in Appendix A.

1.2 Current Methods

Surveillance is typically conducted by personnel trained in the morphological identification of mosquitoes. Not only is it costly to train and employ these personnel, but entomologists are vulnerable to fatigue-based errors in the identification process. For these reasons, mosquito

control districts often forego the surveillance step but still use pesticides (e.g., larvicides and adulticides) to reduce mosquito populations.[13] Unnecessary or otherwise poorly applied insecticide application increases the risk of developing insecticide resistance in mosquito populations.[14]

Older female mosquitoes of pathogen vectoring species are more dangerous in terms of how likely they are to produce an infectious bite resulting in the transmission of an arbovirus to a susceptible human. After acquiring an arbovirus from a blood meal, there is a period before the mosquito host can transmit the arbovirus to a susceptible human where the virus replicates within the mosquito host. The period between acquisition of the arbovirus and the point at which the mosquito host can transmit the arbovirus to a susceptible host is known as the extrinsic incubation period.[15] Dengue, for example, has an extrinsic incubation period of between 8 and 12 days.[16] Determining the age of a mosquito population is important in understanding the risks to human health. A mosquito population with mostly older female mosquitoes that have lived beyond the extrinsic incubation period of a pathogen they vector is clearly more dangerous than a mosquito population with mostly young female mosquitoes that have not lived long enough to be capable of transmitting a pathogen. The rare exception to this general concept is when pathogens are vertically transmitted from the parental generation through transovarial transmission.

Presently, age-grading methods are primarily physiologic and classify female mosquitoes based on parity i.e., parous or nulliparous, where parous mosquitoes have laid eggs while nulliparous mosquitoes have not.[17] This means that a trained entomologist can identify whether or not a mosquito has laid eggs and evaluate a population (or collection) in terms of physiologic age. However, this method does not discriminate the number of gonotrophic cycles or determine

the chronologic age of the mosquito.

1.3 Alternative Methods

Whole cell matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), was used as a molecular protein profiling tool for mosquito species closely related taxonomically.[18] Mosquito heads and thoraces were homogenized into Eppendorf tubes containing formic acid in order to prepare samples. Additionally, a matrix suspension was prepared with sinapic acid in a separate Eppendorf tube.[18] While MALDI-TOF is effective for species identification of closely related mosquito species, the considerable labor coupled with the added expense of consumable reagents makes it unsuitable for use on a large scale. While in theory, this technique could be used to develop methods for chronological age determination of mosquitoes, other techniques exist that are less time consuming and less costly.

A study with *Anopheles gambiae*, the primary malaria vector, used quantitative polymerase chain reaction (qPCR) predict mosquito age. CPR59, which transcribes a cuticular protein, and G12_ANOGA, transcribing a protein G12 precursor, were among the genes involved in the age prediction model.[19] Quantitative PCR is expensive and uses consumable reagents, which serve as barriers for implementation.[20]

With little to no sample preparation, NIR mosquito aging techniques are limited due to the broad spectral bands resulting from the compression and overlapping of many overtone and combination bands.[21] Overtone and combination bands occur when IR light is absorbed and simultaneously excites two or more fundamental transitions.[22] Combination bands are of two or more different fundamental transitions while overtone bands are of the same fundamental transition two or more times.[22] NIR has much lower sensitivity than mid-infrared spectroscopy.[21] This lower sensitivity means that differences in the spectra of different mosquito

samples will be very small and results in lower selectivity than mid-infrared spectroscopy. NIR spectroscopic methods for chronologic age classification, as well as qPCR methods for chronologic age grading, have historically been limited in terms of temporal resolution. These methods of classification often do just that—classify samples into groups of young and old mosquitoes as opposed to predicting age with a linear or non-linear model. [19,23,24]

1.4 Mosquito Background

Mosquitoes may serve as vectors, which means they have the ability to transmit viruses and protozoan parasites (e.g., malaria parasite: *Plasmodium* spp.) between different hosts. While mosquitoes primarily feed on vertebrates, there are rare exceptions. *Unanotaenia sapphirine*, a mosquito of little to no public health importance, is known to utilize annelids as hosts.[25] Pathogen vectoring is possible because many species of mosquito require blood feeding in order to gain the nutrients to produce eggs for reproduction.[26] Not all mosquito species can transmit human pathogens, only about 200 of the approximately 3600 known species of mosquitoes are capable of this.[27]

Female mosquitoes typically become infected when they obtain blood from an infected host. The pathogen must replicate successfully and ultimately replicate in the salivary glands in order for the mosquito to become infectious. Contemporaneously, oogenesis—the development of egg cells into competent cells capable of further development when fertilized—is occurring and the female mosquito will eventually oviposit, or lay, her eggs.

These processes take time, so the gap between subsequent blood meals is often days to weeks. This cycle repeats and eventually female mosquitoes will take a second or third blood meal where pathogen transmission is possible under the correct conditions—the female mosquito has acquired an arbovirus which has replicated over the extrinsic incubation period. Therefore, it

is generally the older female mosquitoes, who have taken multiple blood meals and lived long enough to host an arbovirus through its extrinsic incubation period, that are the most dangerous.

1.4.1 Vectorial Capacity Equation

The vectorial capacity (V) is the total number of potentially infectious bites arising from all the mosquitoes biting a single infectious human on a single day. Vectorial capacity (V) can be defined using the equation:

$$V = \frac{ma^2p^n}{-\ln(p)} \quad (1)$$

where the parasite or virus's extrinsic incubation period is represented by n days, m represents the ratio of mosquitoes to humans, p represents the mosquito survival through one day, and a represents the human biting rates.[28] A human will be subject to the attention of m mosquitoes and will receive bites at the rate of ma^2. For a mosquito to become infectious, they must survive the extrinsic incubation period with the probability p^n where adult mosquitoes live on average $1/(-\ln(p))$ days biting at a rate of a per day. Determining the age of mosquitoes within a population allows a better understanding of the survival rates (p) and a better estimate of vectorial capacity can be made.

1.4.2 *Culex quinquefasciatus* and *Culex tarsalis* background

Culex quinquefasciatus, sometimes referred to as the southern house mosquito, is found in the tropics and warmer temperate regions across the globe. *Culex tarsalis*, sometimes known as the Western encephalitis mosquito, is found across North America most commonly west of the Mississippi River. Both *Culex quinquefasciatus* and *Culex tarsalis* are known vectors for several pathogens including West Nile virus, Western equine encephalitis, and St. Louis encephalitis. These mosquitoes lay their eggs in standing water (natural or man-made) in places such as

swamps or bird baths. These *Culex* mosquitoes are of entomologic interest in the United States, primarily west of the Mississippi River, for their role in human and equine pathogen vectoring.[29]

1.4.3 *Aedes triseriatus* Background

Aedes triseriatus otherwise known as the Eastern tree hole mosquito is commonly found across the eastern half of the United States and Canada.[11] Normally, *Ae. triseriatus* lays its eggs in pools of water that have accumulated such as in tree holes and discarded tires. These mosquitoes generally live in woodland and forested environments and within suburban areas. Eggs from *Ae. triseriatus* can overwinter, utilizing dried containers (natural or artificial) that get flooded with rain in the springtime. While these mosquitoes feed on a variety of non-human vertebrates, they also take blood meals from humans. *Aedes triseriatus* serves as the primary vector for La Crosse encephalitis which can be deadly or cause severe persistent health problems.[11] *Aedes triseriatus* is commonly found in Western North Carolina and understanding more about entomologic risk factors (e.g., arthropod abundance, population dynamics, and infection rates) can greatly improve health measures taken against the spread of La Crosse encephalitis.[30]

1.5 Theory of Infrared Spectroscopy

Infrared spectroscopy uses the infrared region, 12500 cm^{-1} to 10 cm^{-1}, of the electromagnetic spectrum. Within the infrared region, three regions are often distinguished: near-infrared region 12500-4000 cm^{-1}), mid-infrared region (4000-400 cm^{-1}), and far-infrared region (400-10 cm^{-1}). Near-infrared radiation can excite combination and overtone vibrations within molecules, mid-infrared radiation can be used to study fundamental structural vibrations of molecules, and far-infrared radiation can be used to study vibrations of heavy atoms or large-scale vibrations. Since most molecules have characteristic absorbances and primary molecular vibrations within the mid-infrared region, it is the most commonly used for analysis.[31]

Spectroscopy is the study of the interaction between light and matter. When infrared light interacts with different groups of atoms in molecules, photons of specific wavelengths are absorbed, which excites the groups to higher energy states. A molecule must undergo a net change in dipole moment as it vibrates in order for infrared absorption to occur. The maximum number of vibrations (n) or modes within a molecule can be calculated using the following equation for nonlinear molecules:

$$n = 3N - 6 \tag{2}$$

where N represents the number of atoms in a given molecular structure. When absorption occurs, a vibrational transition from the ground state ($v = 0$) to the first excited state ($v = 1$) occurs, where the gap between the energy levels (ΔE) corresponds to the frequency of light that excited the molecule and can be calculated with the equation:

$$\Delta E = hc\tilde{\nu} \tag{3}$$

where h is Planck's constant, c is the speed of light, and $\tilde{\nu}$ is the wavenumber of the light.

When a molecule vibrates, there is a change in the net dipole moment related to the orientation of the electrons in the molecular electric field.[32] Each molecule will have a unique infrared spectrum where bands correspond to different vibrations or combinations of vibrations within the molecule. The wavenumber for different vibrational modes can be described by the equation:

$$\tilde{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \tag{4}$$

where μ is the reduced mass of the atoms involved in the vibration and k is the spring constant that represents the strength of the bonds involved in the vibration. Keeping in mind that wavenumber is directly proportional to energy, using equation 4, vibrations involving larger

atoms with weaker bonds such as single bonds between a carbon atom and a hydrogen atom, will have lower energy vibrational wavenumbers. Conversely, smaller atoms with stronger bonds, such as double or triple bonds, will have higher energy vibrational frequencies. Moreover, there are direct connections between the bands in an infrared spectrum and the structure of the molecule.[32]

1.6 Infrared Spectroscopy Sampling Techniques

Mid-infrared spectroscopy is a well-established and reliable technique for chemical fingerprinting that can obtain spectra from a wide variety of samples of solids, liquids, and gases. Each molecule will have a unique pattern of peaks which serves as a chemical fingerprint. The way by which the sample is handled can have a large effect on the consistency of a measurement. Developing a method for sample handling is paramount in making consistent measurements while returning the best quality of spectra. For infrared spectroscopy there are several sample handling techniques commonly used: transmission, attenuated total reflectance (ATR), specular reflectance, and diffuse reflectance.

In the transmission technique for infrared spectroscopy, the IR beam passes through the sample where some of the IR light interacts with the sample, the light that doesn't interact with the sample reaches the detector, and a spectrum is generated based on the energy of the transmitted light. This technique has a high signal to noise ratio and historical prevalence. However, transmission mid-infrared spectroscopy requires expertise in sample preparation which is both time consuming and difficult to reproduce.[33]

Attenuated total reflectance (ATR) is a surface sampling technique often combined with infrared spectroscopy for liquid and solid-state samples. This technique requires little to no sample preparation. The sample is loaded onto an optically dense crystal with a high refractive

index, such as a diamond, and the infrared light is directed onto the crystal at an angle. ATR utilizes the optical phenomenon of total internal reflection which results in an evanescent wave that penetrates into a sample between 0.5 and 2 μm. While ATR is extremely robust and requires little to no sample preparation, it is often destructive to the sample or leaves it otherwise unrecoverable.[33,34] Any air between the sample and the crystal can affect spectral data so pressure must be applied to the sample to minimize this effect.[33]

Specular reflection is the mirror-like reflection of light on the surface of a sample where the incidence angle and angle of reflection are the same but are on opposite sides of the surface normal—perpendicular to the sample surface at the point of contact between the infrared source and the sample. This technique is often used to provide qualitative data about a surface such as a thin layer polymer or a coating on a polished metal.[33] For this technique to be effectively used, the sample must be large, flat, and have a reflective surface.

Diffuse reflectance or DRIFTS is a commonly used reflectance measurement technique where the incidence angle and reflected angle are not necessarily the same, such as on rough surfaces. This sample preparation technique requires little to no sample preparation and occurs more frequently in everyday environments than specular reflection.[33,35]

1.6.1 Infrared Microspectroscopy

While many infrared spectroscopy sampling techniques can be used for chemical identification, combining them with microscopy (microspectroscopy) allows the deciphering of complex and spatially heterogeneous samples. A reflectance microspectrometer will utilize the reflective properties of a shiny metal substrate upon which the sample is placed. Infrared light travels to the sample and the reflective properties of the sample and the shiny metal substrate return that light to the detector. With reflectance microspectroscopy both specular and diffuse reflection are

utilized to return information about the chemistry and the topography of the sample in the resulting spectrum.[31]

1.6.2 Mid-Infrared Spectroscopy for Age Prediction

Mid-infrared spectroscopy is a type of vibrational spectroscopy that is both sensitive and information rich. This type of spectroscopy is capable of measuring a wide variety of molecular signals ranging from transmembrane protein-lipid interactions to subtle changes in protein secondary structure.[36-40] Infrared spectra have spectral features that can be used to distinguish between or identify different biological samples. The combination with numerical analysis or multivariate data analysis allows information relevant to classification to be extrapolated from high dimensional spectral data. This broadens the capability of mid-infrared spectroscopy in addition to being fast.[41]

Mid-infrared spectroscopy has already been used to characterize and identify microorganisms, viruses, and types of cancers.[42,43] Vibrational spectroscopy has been used to identify and classify different species of bacteria especially foodborne pathogens.[38,42] As a bio-analytical tool, vibrational spectroscopy has been used to differentiate between normal and pathological tissue including numerous types of cancers: breast, endometrial, cervical, prostatic and brain cancers.[37]

While mid-infrared spectroscopy is a well-established and information rich technique, its applications in the mosquito world for classification and identification are limited. Mid-infrared based techniques were recently utilized to detect *Wolbachia* infection in *Aedes aegypti* mosquitoes in Australia.[36] These techniques were also used to identify sex and distinguish between two age groups of *Ae. aegypti* (2 and 10 days).[36] Additionally, mid-infrared microspectroscopy coupled with advanced numerical analysis was employed to classify four

species of *Aedes* mosquitoes.[3] The use of mid-infrared microspectroscopy in chronologic age grading has not otherwise been well established.

1.7 Data Processing Theory

Data processing is broken into two steps: training and validation. After sample preparation, spectral acquisition, and quality control steps, the data is randomly divided in half. One half of the data (training) is used to develop a model to classify the other half (validation). The validation set predictions are used to establish the accuracy and uncertainty in the model's performance.

Within data analysis there is data pre-processing which bolsters the robustness and accuracy of the subsequent numerical analyses in addition to correcting for variance in the data acquisition step. Due to the variation in the shape of biological samples and a constantly changing background due to the microspectrometer being open to the atmosphere, a number of inconsistencies arise when acquiring infrared spectrum of these samples. Most commonly, the baselines of spectra can be slanted or oscillatory due to scattering, interference from the sample shape, interference from carbon dioxide and water vapor, and variation in sample thickness.

1.7.1 Cropping

The goal of cropping the data is to remove portions of the spectra where atmospheric contributions to the signal, such as water vapor or carbon dioxide, are prevalent.[31] Concurrently, areas of the spectra that have little to no chemical information can be removed as well. Cropping of the data can result in the loss of significant chemical information or in the creation of artifacts at the edges of the spectral data. The data are cropped to an information rich region of the infrared spectrum where there is the greatest likelihood of interpretable differences in the spectra of different samples.

1.7.2 Normalization

Normalization is intended to account for the variance of sample thickness. There are many different ways to normalize the data, but normalization by height assigns the values for the spectra between 0 at the lowest band and 1 for the highest band.

1.7.3 Savitzky-Golay Smoothing and Second Derivative Function

Baseline interference is very common in spectra of biological samples due to scattering caused by the shape of the sample. Additionally, low frequency instrument noise can influence the baseline of the spectra. To correct for these issues, a Savitzky-Golay smoothing and second derivative function are applied to the spectra.[44-46] This technique aims to minimize the differences in spectra caused by noise or baseline oscillation while amplifying the differences in the chemical information of the spectra.

1.7.4 Mean-Centering

Mean centering is employed to reduce redundancies and simplify the data which might help in classification. This technique subtracts the average spectra from all the spectra. Redundancies are subtracted out from the spectra while differences from the mean data remain and are enhanced comparatively. Mean centering the data decreases the complexity of the data and reduces the number of factors required to model the data.[47]

1.8 Data Analysis Tools

1.8.1 Principal Components Analysis

Principal components analysis (PCA) is an exploratory data analysis tool. PCA does not make any assumptions about the data. This tool results in loading vectors, or principal components, which contain composite spectral information. Each loading vector is orthogonal to the next meaning that there is ideally no overlap of the information between the loading vectors. Scores

are the weights of the loading vectors and the product of the scores and the loading vectors returns the original spectrum:

$$A = S\,L \tag{5}$$

where A is the original spectrum, S represents the scores, and L represents the loading vectors. Scores plots are how the data are explored. The score values serve as coordinates on a Cartesian Plane where each axis e.g., x, y, z, is a different score. Groups of the points in the scores plot can be color coded to identify the groups. Figure 1 shows an example of a scores plot where the values for score Y versus score X are plotted. Red points represent young (< 1 week old) mosquito spectra while blue points represent old (≥ 2 weeks old) mosquito spectra.

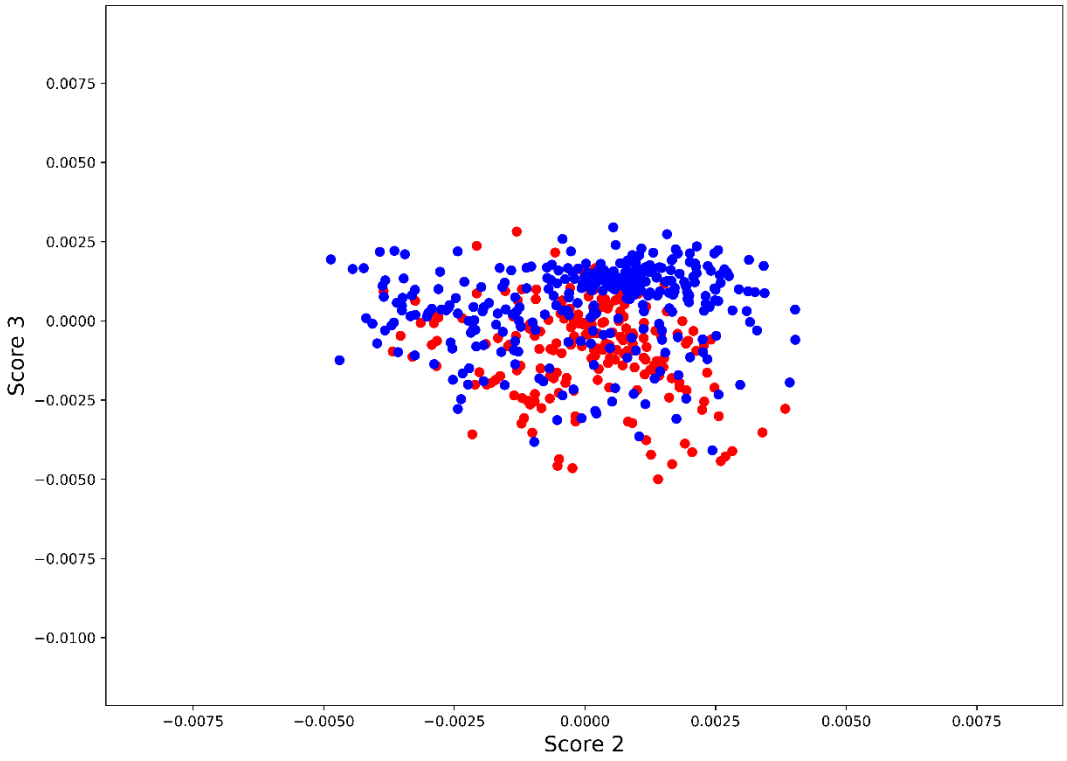

Figure 1. Principal components analysis scores plot example where score 3 values are on the y-axis while score 2 values are on the x-axis. Red points represent young (< 1 week old) mosquito spectra and blue points represent old (≥ 2 weeks old) mosquito spectra.

1.8.2 Principal Components Analysis Regression

Principal components analysis regression or PCAR is a regression analysis technique based on PCA. PCAR is used to estimate unknown regression coefficients by using the principal components of the explanatory variables as regressors instead of regressing the dependent variable on the explanatory variable directly.[48] Only a subset of the principal components are used in this type of regression analysis. It is important to select those principal components which are important for prediction. PCAR is often used to overcome multicollinearity where two or more of the explanatory variables are close to being linear. This is achieved by limiting the number of principal components and excluding some low-variance principal components in the regression step. While only using a subset of the principal components, the number of parameters characterizing the model is reduced resulting in a dimensional reduction, limiting the number of random variables influencing the model. If the principal components are selected appropriately, PCAR can be effectively used to predict outcomes based on a model dataset.

1.8.3 Partial Least Squares Regression

Partial least squares regression (PLSR) is a method for relating two data matrices by a linear or near linear multivariate model. Partial least squares use latent variables, or variables not directly observed but inferred by other variables, to model the covariance of two matrices.[49,50] Like PCA, PLSR uses principal components, or loading vectors, but they are calculated differently. In PCA principal components are found by maximizing the spectral variance while PLSR aims to maximize the covariance of the independent and dependent variables. PLSR models work best when the predicting matrix has more variables than observations and when multicollinearity exists among the predictor matrix values.[49,50]

1.8.4 Artificial Neural Networks

An artificial neural network (ANN) can be defined as sophisticated nonlinear computational tools, which are capable of modeling functions of extreme complexity.[51] Using an appropriate ANN architecture can represent nearly any functional relation between a set of inputs and outputs. A neuron serves as the fundamental unit where a single calculation takes place usually with multiple inputs and a single output.[51] The simplest way to represent an ANN is with a black box that receives multiple inputs and produces multiple outputs.[51] The black box scheme for ANN is represented in Figure 2 where x_1, x_2, x_3, …, x_m are the input variables (e.g., spectra or spectral information) and y_1, y_2, …, y_n are the output variables (e.g., predictions of mosquito age) after processing in the black box.[51]

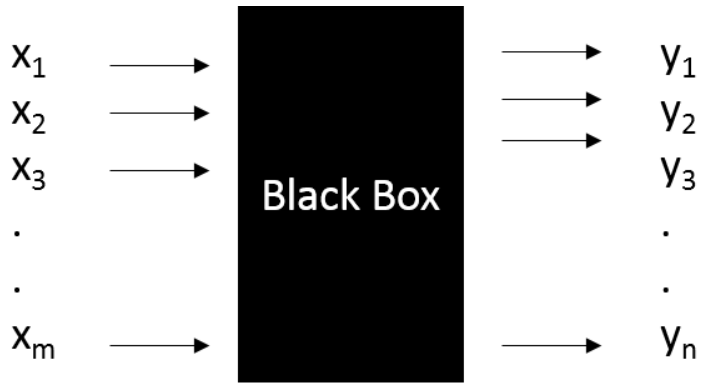

Figure 2. Black-box scheme of artificial neural network comprised of input variables (x_1, x_2, x_3, …, x_m), the black box, and output variables (y_1, y_2, …, y_n).

Neurons are often organized in layers where neurons within the same layer are operating on inputs of the previous layer. The input layer passes input information to the hidden layer as variables.[51] Hidden layers perform computations on the input variables while the output layer performs the final calculation.[51] Each neuron of the input layer passes information to each neuron of the hidden layer which passes information to each output neuron. This type of ANN

architecture is called a multilayer perceptron, or multilayer feed-forward, neural network and is represented in Figure 3 with an input layer (x_1, x_2, …, x_5), a hidden layer (h_1, h_2), an output layer (y_1, y_2, …, y_4), and arrows connecting the neurons in the layers showing how information is passed. Each calculation can have a different weight associated with it in order to achieve a different output. By training an ANN model, different weights of calculations are used to best connect the input variables with the output variables. This process is usually done with a subset of the data known as a training dataset. Once the artificial neural network has 'learned' or developed a connection between the input and output variables using the training dataset, the model can be tested on new data to evaluate the model's performance, usually the validation dataset where predicted values can be compared to true values.

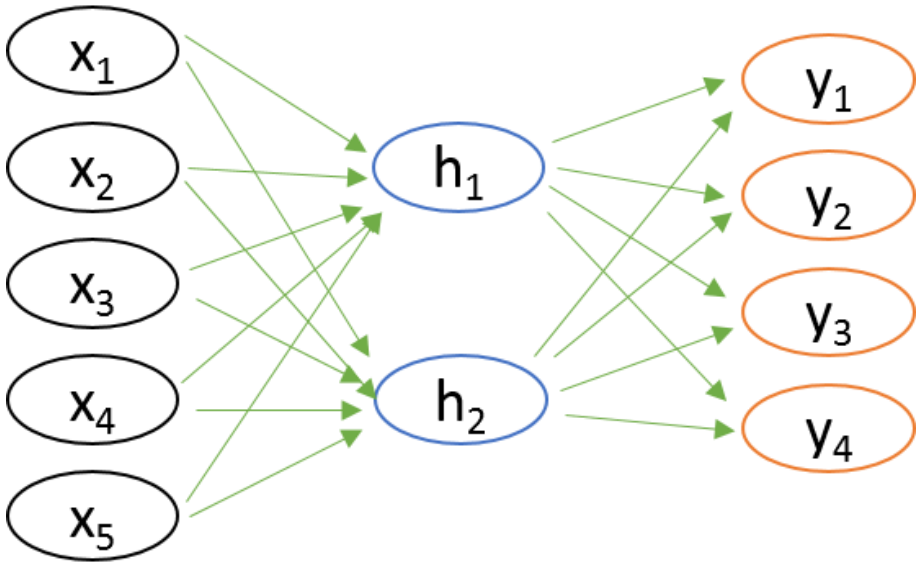

Figure 3. Multilayer feed-forward artificial neural network where the input layer (x_1, x_2, …, x_5) passes information to the hidden layer (h_1, h_2) where calculations are performed. Information from the hidden layer is passed to the output layer (y_1, y_2, …, y_4) where the final calculations are made. The arrows represent the passing of information from one neuron to another.

1.9 Research AIMS and Hypotheses

AIMS: Discriminate categorical and quantitative differences in ages of *Culex quinquefasciatus, Culex tarsalis,* and *Aedes triseriatus* using mid-infrared spectroscopy and chemometrics. Create a model using mid-infrared spectroscopy and chemometrics to predict the chronologic age of *Aedes triseriatus*. Investigate the biochemical differences that change as a function of mosquito age.

Hypotheses: There are biochemical differences in mosquitoes of different ages that can be detected using mid-infrared spectroscopy and chemometrics and those biochemical differences can be used to build PLSR and ANN models for chronologic age prediction of *Aedes triseriatus* in a time series study.

CHAPTER TWO: EXPERIMENTAL

2.1 Materials & Methods

2.1.1 Rearing Materials & Methods

Culex quinquefasciatus strain JHB, were obtained from MR4/BEI resources. The colony was initially established from field collected samples collected at a pond north of Johannesburg, South Africa (Coordinates 26° 66'S 27° 50'E). The colony was contributed to MR4/BEI by A.J. Cornel. *Culex tarsalis* (strain YOLO) were obtained through BEI Resources, (NIAID, NIH: *Culex tarsalis* YOLO, NR-43026). The specimens were originally sourced from a dry ice bait trap, Fazio Wildlife Refuge, Yolo County, CA in 2003. *Aedes triseriatus* (strain MSU) were obtained from Michael Kaufman at Michigan State University in 2018. The MSU strain is kept in continuous colony at WCU (WCU Mosquito and Vector-borne Infectious Disease Laboratory) and has an unknown generation history.

A 5% yeast solution in water (v/v) was used to initiate hatching while a solution of 5% liver powder in water (v/v) was used to feed the larvae as they transitioned through different instar or developmental stages. Once mosquito larvae pupate, i.e., transition from larvae to pupae, they no longer feed as they use the nutrients acquired as larvae to develop into adult mosquitoes. Both the 5% yeast solution and 5% liver powder solution were added to trays filled with water and unhatched mosquito eggs. Liver powder solution was added every few days such that there were visible food particles on the bottom of the tray. Mosquitoes transition through 4 different instar stages where they molt (shed) their exoskeleton on each occasion growing in size until they metamorphose into pupae after the fourth molt.[52] Mosquitoes will remain at this stage for only a few days, so the pupa were manually transferred using a plastic transfer pipet to pupal

rearing chambers. Pupal rearing chambers are bi-sectional enclosed structures which allow eclosion (i.e. the transition from pupae to adult) and for feeding and sustaining adult mosquitoes. The sections are separated by a funnel such that mosquitoes can fly into the upper chamber, but it is difficult to fly back into the lower chamber. In the upper chamber, the mosquitoes can feed on cotton ball soaked in a sucrose solution through mesh at the top end of the container made with 5% (v/v) Karo® Light Corn Syrup in water. Figure 4 shows an example of a pupal rearing chamber. Pupae are able to swim in the water at the bottom of the lower chamber while adults live in the space above the water and feed through mesh at the top of the chamber.

Once these mosquitoes reached adulthood, the adults were generally held in a Percival i41vl incubator (Perry, Iowa) at 27°C and 70% RH. During one experiment (*Ae. triseriatus* time-series), WCU was at reduced operations due to the COVID-19 pandemic.[53] In anticipation of limited access to the laboratory, the *Ae. triseriatus* cohorts were moved to a private residence where the cage environmental conditions differed from the laboratory: temperature range: 19°C-22°C, and RH range: 50-65%. Overall, for this particular experiment, the temperature range was 19°C-27°C and 50-75% RH. Mosquito samples were killed by freezing at different ages in order to produce cohorts (age bins) of mosquitoes at different ages.

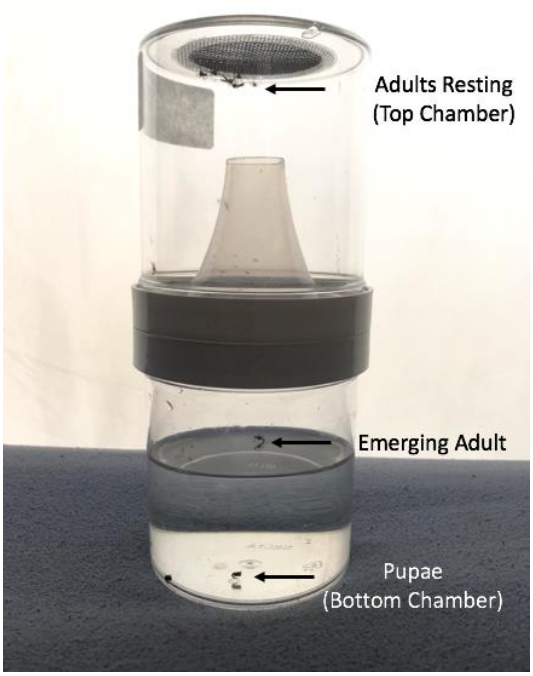

Figure 4. Pupal rearing chamber where pupae are able to metamorphose in the water in the bottom chamber and adults can fly around in the remaining air in the chamber. Adult mosquitoes feed through mesh at the top of the pupal rearing chamber.

2.1.2 Mosquito Samples

A total of 280 male and female *Culex quinquefasciatus* samples were acquired and stored in a freezer (-80 °C) before and after measurement to preserve the biochemistry of the samples. These samples were divided into two bins: young mosquitoes < 1 week old (n = 105), and old mosquitoes ≥ 2-weeks old (n = 156) based on initial holding times.

Approximately 80 total male and female *Culex tarsalis* samples were measured and stored in a freezer (-80 °C). *Culex tarsalis* samples were divided into 5-day age bins where the ages of the mosquitoes in each bin were ±2 days (e.g., 5-day bin mosquito ages ranged from 3-7 days old). A total of 279 female *Aedes triseriatus* samples were acquired and stored in a freezer (-80 °C) of which 210 were measured. *Aedes triseriatus* samples were divided into age bins (n = 30) of 1 day (20-24 hrs), 2 days (>24 hrs, < 30 hrs), 7 days, 14 days, 21 days, 28 days, and 35

days. The mosquito species, age bins, and the uncertainty in age used in this study can be found in Table 1.

Table 1. Mosquito sample age distribution.

Species	Bins	Uncertainty in Age
Culex quinquefasciatus	< 1 week (n = 105) ≥ 2 weeks (n = 156)	$\Delta t \geq 7$ days
Culex tarsalis	5 days (n = 11) 10 days (n = 11) 15 days (n = 18) 20 days (n = 17) 25 days (n = 12) 30 days (n = 8) 40 days (n = 4)	$\Delta t \leq 2$ days; e.g., 5 days ±2 days
Aedes triseriatus	1 day (20-24 hrs) (n = 30) 2 days (>24 hrs, < 30 hrs) (n = 30) 7 days (n = 30) 14 days (n = 30) 21 days (n = 30) 28 days (n = 30) 35 days (n = 30)	$\Delta t = \pm 1$ day

2.1.3 Instrumentation

Specimens were measured using a ThermoNicolet™ model Centaurus IR microscope with a liquid nitrogen cooled mercury cadmium telluride (MCT/A) detector attached to a ThermoNicolet™ iS10 bench, a photo of which can be found in Figure 5 (WCU, Cullowhee, NC, USA). The OMNIC™ version 9.8.372 software was used to collect all spectroscopic data. Backgrounds were collected roughly every 10 minutes to 15 minutes. The interval for background collection is dependent upon the subjective judgment of the user. The obvious presence of water vapor between 2000 cm^{-1} and 1800 cm^{-1} was one of such indicators. All spectra were acquired from the sample at room temperature (20-23 °C). Samples were stored in a freezer (-80 °C). A full list of instrumentation parameters is listed in Table 2.

Figure 5. Photo of ThermoNicolet™ model Centaurus IR microscope used to make all spectroscopic measurements.

Table 2. Instrumentation parameters for ThermoNicolet™ model Centaurus IR microscope used to make all spectroscopic measurements.

Parameter	**Value**
Microscope Make & Model	ThermoNicolet™ model Centaurus
Bench Make & Model	ThermoNicolet™ IS10
Software	OMNIC™ version 9.8.372
Wavelength Range	4000 cm^{-1} – 650 cm^{-1}
Detector	MCT/A, Liquid Nitrogen Cooled
Beamsplitter	KBr
Scans	64
Blank	Gold Microscope Slide in Air

2.1.4 Measurement Procedure

Samples were prepared by using a Lecia L2 stereomicroscope (which uses series of lenses to magnify the sample and lights which reflect off of a mirror to illuminate the sample) to remove one of the hind legs, or to identify sample sex if necessary. A photo of the Lecia L2 stereomicroscope can be found in Figure 6 (WCU, Cullowhee, NC, USA). The ThermoNicolet™ model Centaurus infrared microspectrometer was used to measure all mosquito samples. In an infrared microspectrometer, the IR radiation passes through the upper objective and projects IR radiation onto the sample. Infrared radiation will penetrate the leg in addition to reflecting off of the surface in the form of specular and diffuse radiation before returning to the objective and the detector. Within the infrared microscope is a camera that allowed easier focusing of the IR radiation onto the sample leading to better reproducibility. A clean gold-coated microscope slide was measured as the background, and samples were placed on a shiny metal plate on the microscope stage when measuring. The middle of the tibia of the hind leg was measured for each sample. An internal standard was used for the *Ae. triseriatus* where a black sharpie mark directly on the gold-coated microscope slide was measured at the beginning and end of measuring and roughly once an hour between. Once sample collection was complete, files were saved as SPA and JDX file formats. The SPA file format is used by the OMNIC™ software and allows spectra to be viewed and compared at the instrument. The JDX files were compiled into an HDF or hierarchical data format file which allows for easy transfer of large data sets.[54] This HDF file was transferred to a personal computer where data processing was completed.

Figure 6. Leica L2 Stereomicroscope used for sample preparation and sex identification. Under the stage is a mirror that reflects the light to the bottom of the sample.

2.1.5 Exploratory Data Analysis

Exploratory analysis was conducted in order to remove outliers resulting from spectral acquisition. Spectra with excessive oscillation in the baseline were noted as the intensities of some bands containing chemical information would likely be skewed higher or lower. Excessive oscillation of the baseline contributions usually resulted from samples not sitting flat on the shiny metal plate during the measurement. When a sample was not sitting flat on the metal plate, it could sometimes be corrected, and the measurement repeated for that sample. Other times, baseline oscillation was likely due to a twist in the mosquito leg that could not easily be corrected. Moreover, exploratory data analysis mitigated some of the non-chemical differences between samples in order to produce a better predictive model.

2.2 Data Processing

2.2.1 Data Pre-Processing

A number of data pre-processing steps were taken in order to minimize variance from the spectral acquisition process and to enhance the differences between samples that might be used for age prediction.

Spectra were cropped to an information rich region between 1800 and 650 cm^{-1} or 1800 and 1000 cm^{-1} depending on the experiment. Spectra were normalized by band height where the highest peak for each spectrum was set to a height of 1 and the remaining band heights were between 0 and 1 relative to the highest peak. A second derivative Savitzky-Golay algorithm with a 25 cm^{-1} window size and a second-degree polynomial were applied to the spectra. For the PLSR model, data were mean centered by subtracting the mean spectra from each of the spectra in order to mitigate redundancies and emphasize differences in the spectra.

After data pre-processing steps, spectra look much different and are no longer directly interpretable in terms of the chemical information corresponding to each band. Figure 7 shows *Aedes triseriatus* spectra following pre-processing with cropping to 1800 – 1000 cm^{-1}, normalizing by band height, applying a Savitzky-Golay smoothing algorithm, and applying a second derivative function with the second derivative of absorbance on the y-axis and wavenumber on the x-axis.

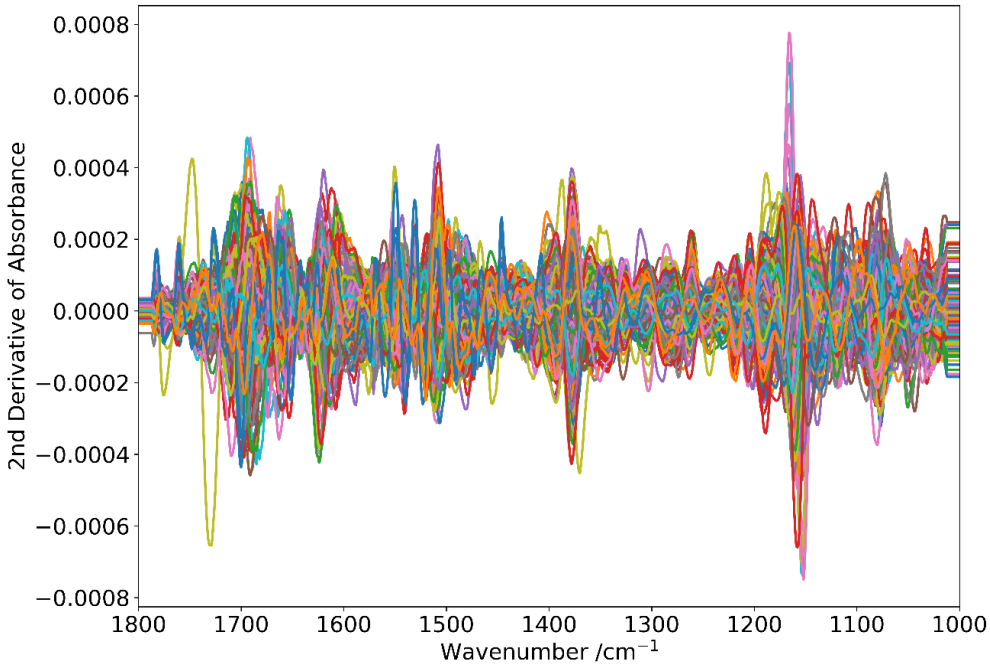

Figure 7. Overlay of female *Ae. triseriatus* spectra after cropping to 1800 –1000 cm^{-1}, normalizing by band height, applying Savitzky-Golay smoothing & 2nd derivative function, and mean-centering.

2.2.2 Principal Components Analysis

Principal components analysis was used primarily for qualitative analysis to visualize the grouping of mosquito samples of different ages. After pre-processing, data were loaded into ten loading vectors. Score plots were used to visualize groups of old and young mosquitoes with different colors for the different groups. Separation between groups was optimized by hand using qualitative analysis of the grouping with different scores. For *Cx. quinquefasciatus* scores 2 and 3 were used to visualize the separation between 'young' mosquitoes and 'old' mosquitoes. For *Cx. tarsalis* scores 4 and 3 were used to visualize the separation between age bins 5 & 10 and 15, 20, 25, 30, & 40.

2.2.3 Principal Components Analysis Regression

Principal components analysis regression was used to help find outliers in the *Ae. triseriatus* data

after spectral acquisition and data pre-processing. Data were randomly split into training and validation data sets where 4 loading vectors were used to fit the mean training data with age. The true age values were extracted, and residuals of true age and predicted age were calculated. These residuals were plotted and spectra with large differences between predicted and actual ages were inspected and dropped from the dataset if deemed low in quality. Low quality spectra dropped from the dataset primarily had issues with an oscillatory or slanted baseline or had issues with water vapor interference or peak shape.

2.2.4 Model Evaluation

To evaluate the performance for a model, the standard error of prediction (SEPsv) was calculated using the following equation:

$$\text{SEPsv} = \sqrt{\frac{\sum(residuals^2)}{n}} \tag{6}$$

where the difference between the predicted and actual ages, or $residuals$, were squared for each sample predicted. The sum of the square of the $residuals$ was divided by the number of spectra (n). Finally, the square root was taken to give the standard error of prediction. A separate validation scheme was chosen (i.e., training and validation sets) because the number of independent variables (age) was not much less than the number of dependent variables (spectra).

2.2.5 Model Optimization

Training was used to further optimize the predictive performance of the models. The training dataset was further split into training and validation datasets where parameters were optimized for predicting with this subset of the training data. The optimized model was used to predict the validation dataset where the SEPsv was calculated.

2.2.6 Partial Least Squares Regression

Partial least squares regression was used to create a model for predicting the age of mosquitoes

based on their spectra. Data were optimized as described above and 5 loading vectors were used in this model for prediction.

2.2.7 Artificial Neural Networks

A multilayer perceptron neural network was used to create a model for predicting the age of mosquitoes based on their spectra. Data were sorted into training and validation datasets as described above. This model was optimized by hand and four hidden layers of sized 100, 75, 50, and 25 were used.

2.2.8 Age Prediction Workflow

Figure 8 describes the workflow used for age prediction starting with sample preparation and acquisition followed by data pre-processing and splitting of the dataset. The data were split into training and validation datasets where the training dataset was used to develop the predictive model. In order to optimize the predictive model, the training dataset was broken into a further training and validation dataset. After an optimized set of parameters was found for the predictive model, the model was used to predict the ages of the mosquitoes in the validation dataset where the performance of the models were evaluated using the standard error of prediction.

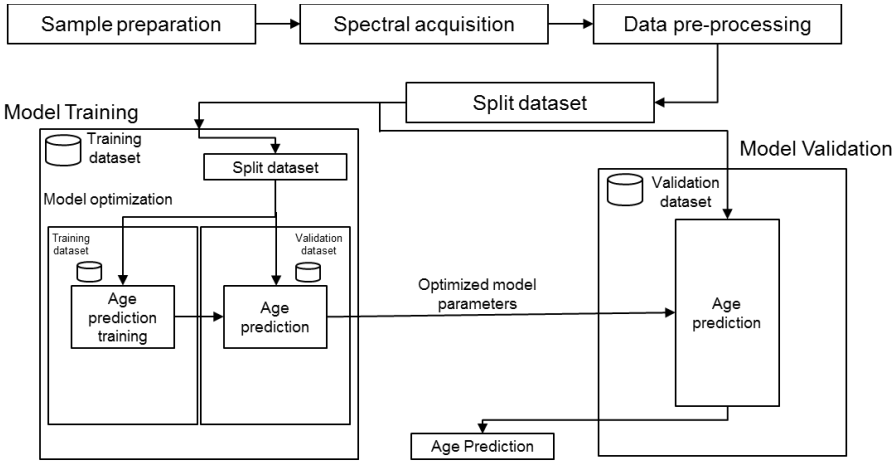

Figure 8. Workflow for age prediction with separate training and validation datasets.

CHAPTER THREE: RESULTS AND DISCUSSION

3.1 Study 1: Age-grading *Culex quinquefasciatus* (old vs. young)

Culex quinquefasciatus were killed off by freezing when they were <1 week in age (young). The second population was killed off by freezing when they were >2 weeks of age (old). A total of 280 *Cx. quinquefasciatus* samples were measured. Of the 280 samples measured, only spectra from female samples were used to calculate principal components and scores for PCA (n = 132) as female mosquitoes are relevant to epidemiology whereas male mosquitoes are less so. All samples were prepared and measured as described in section 2.1.4. In Figure 9 the average spectra for young and old mosquitoes is shown with young mosquitoes, less than one week of age, represented in orange and old mosquitoes, at least two weeks of age, represented in blue. The difference in relative heights of bands indicates a spectral, and therefore chemical, difference between young and old mosquitoes. Principal components analysis was performed on the pre-processed data and the scores plot of score 3 versus score 2 is shown in Figure 10 where red points represent spectra of young mosquitoes and blue points represent spectra of old mosquitoes. Scores 3 and 2 were chosen for the scores plot as this combination offered the best separation between groups of young and old *Cx. quinquefasciatus* mosquitoes.

This experiment served primarily as a proof of concept—that mid-infrared spectroscopy and chemometrics could be used to differentiate between two different age groups of mosquitoes. While pre-processing steps helped to mitigate the differences in mosquito spectra that occurred during spectral acquisition, greater separation between the groups of young and old mosquitoes could likely have been achieved by investigating potential outliers and dropping them from the dataset such as the point in the upper left corner of Figure 10.

Figure 9. Average mid-IR spectra of *Cx. quinquefasciatus* < 1 week old (orange) & ≥ 2 weeks old (blue).

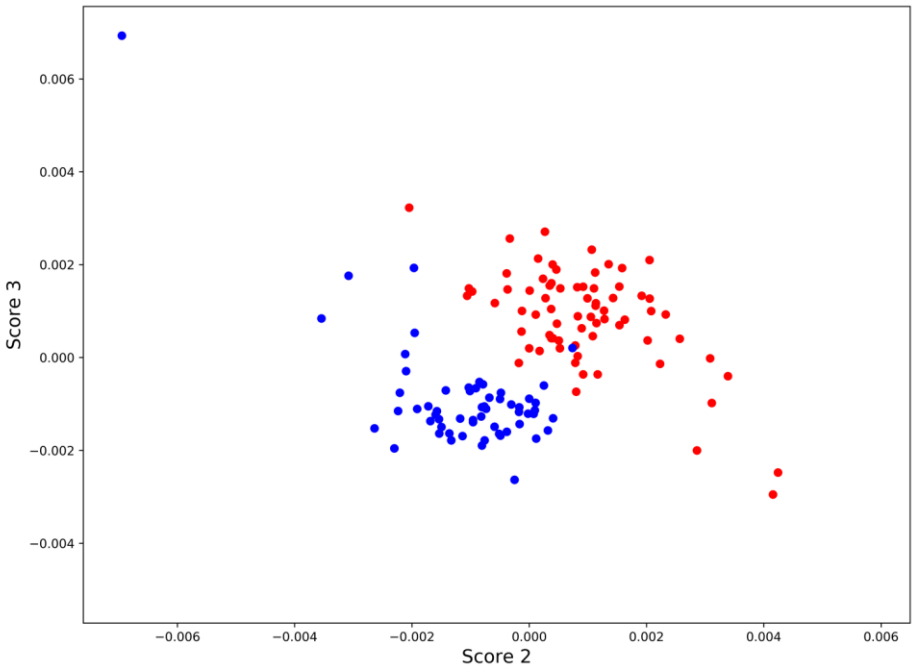

Figure 10. Principal components analysis score plot of female *Cx. quinquefasciatus* where the red points are young (< 1 week old) samples and the blue points are old (≥ 2 weeks old) samples.

3.2 Study 2: Age-grading *Culex tarsalis* (old vs. young: 5-day age bins)

Following the success of separating *Cx. quinquefasciatus* into groups based on age, a new experiment was designed in order to obtain samples of mosquitoes at numerous ages—age bins of mosquitoes. Adult mosquitoes were killed off via freezing at ages of 5, 10, 15, 20, 25, 30, and 40 days (Table 1). A total of 81 male and female *Cx. tarsalis* were measured of which 31 were female. As discussed in Section 1.4, female mosquitoes are of more epidemiologic significance than males, so the analysis of those spectra were prioritized and discussed below. Principal components analysis was performed on the pre-processed data and the scores plot of score 3 versus score 4 is shown in Figure 11 where red points represent spectra of mosquitoes in the 5 & 10 day age bins and blue points represent spectra of mosquitoes in the 15, 20, 25, 30, & 40 day age bins. Scores 3 and 4 were chosen because this combination of scores showed the most separation between the 5 & 10 day age bins from the 15, 20, 25, 30, & 40 day age bins.

While primarily designed to serve as a practice time series analysis to produce mosquito samples of varying ages, this experiment also reaffirmed that mosquitoes could be separated by age using mid-infrared spectroscopy and chemometrics. The time series study served to identify potential variables and such as the insemination of female mosquitoes by male mosquitoes in the same colony space, the uncertainty in mosquito age within each group, and groups sizes for each age bin. Additionally, since female mosquitoes are those most relevant for epidemiology, it did not make sense to continue measuring male mosquitoes. While the true distribution of age for mosquitoes in a time course study is continuous, the grouping into age bins turns age into categorical data. By decreasing the uncertainty, each category is truer to the age it represents and might be more likely to result in a better predictive model. Moreover, increasing group size would encompass more variability within a mosquito population for a given age and also help to

strengthen a predictive model.

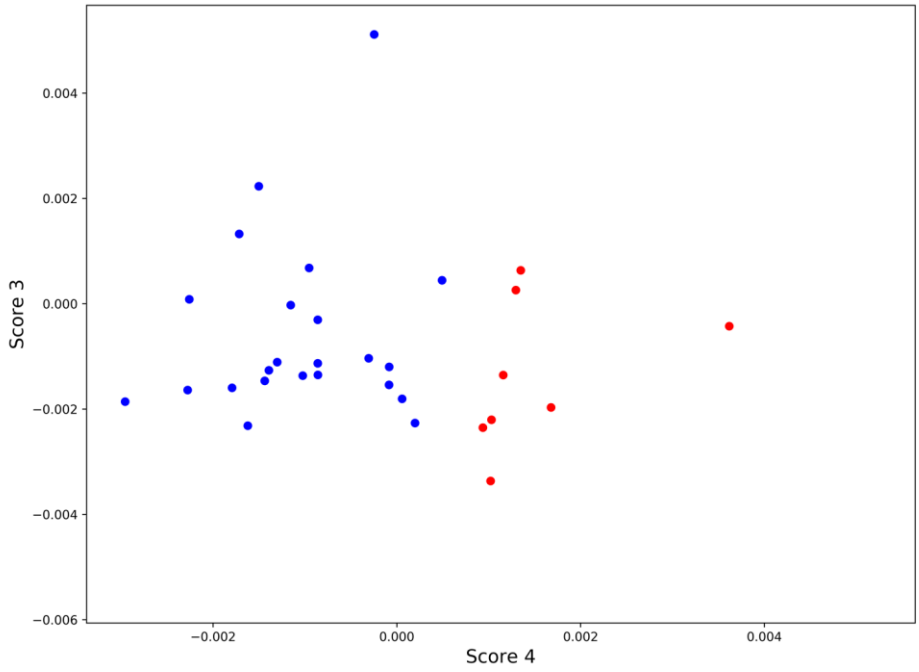

Figure 11. Score plot of female *Cx. tarsalis* samples for 5 & 10 day bins in (red) and 15, 20, 25, 30, & 40 day bins (blue).

3.3 Study 3: Age-grading *Aedes triseriatus* (1 week age bins)

Aedes triseriatus were reared and separated into age bins in a manner similar to as described in section 3.2 with less uncertainty in age. Due to circumstances related to the COVID-19 pandemic, the last four groups (ages 14, 21, 28, & 35) were maintained at a broader relative humidity and temperature range (see section 2.1.1).[53]

A total of 210 female *Ae. triseriatus* were measured using the sample acquisition procedure described in section 2.1.4. After dropping outliers, remaining spectra (n = 162) were cropped to the region of 1800 – 650 cm^{-1} to inspect for any highly variable regions in the data as shown in Figure 12. Another method for looking at the spectra plotted each spectra by age group shown in Figure 13. Spectra were pre-processed by cropping to 1800 – 1000 cm^{-1}, normalizing

by band height, applying a Savitzky-Golay smoothing algorithm, and applying a second derivative function. The data were mean centered for the PLSR model but were not for the ANN model.

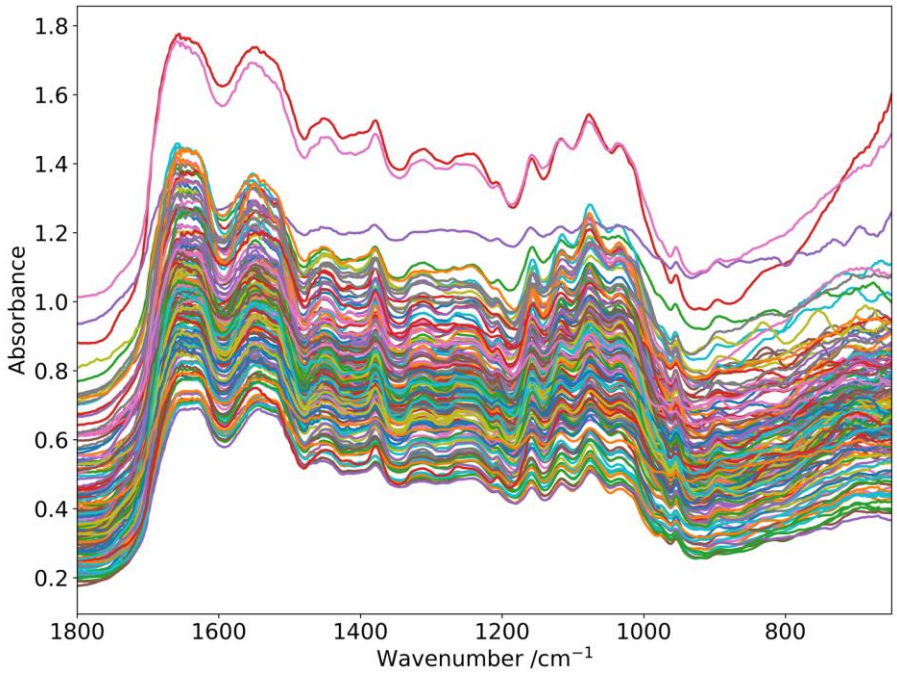

Figure 12. Overlay of 162 female *Ae. triseriatus* spectra cropped to the region of 1800 to 650 cm^{-1}.

The optimized partial least squares regression model resulted in a standard error of prediction (SEPsv) of 4.3 for the validation set of mosquitoes. A plot of the predicted age versus known age is shown in Figure 14. Performance of the PLSR model can be qualitatively assessed by looking at the vertical distribution of points for each age group; the smaller the vertical distribution, the higher the performance. A positive linear trend of points can be observed except for between groups 1 & 2 where the trend is reversed.

The optimized artificial neural network model resulted in an SEPsv of 3.3 for the validation set of mosquitoes was also used to create a model for predicting the age of mosquitoes

using a training dataset. Predicted age by the ANN model versus actual age was plotted and is shown in Figure 14. A linear trend can be observed that predicted the ages of 1 & 2 days much better than the other days. The performance of the ANN model can be qualitatively assessed using the vertical distribution of points at each known age; the smaller the vertical distribution, the better the performance. Overlaid, it is easy to compare the performance of both models and see that ANN predicted the ages of groups 1, 2, and 7 better when compared to the PLSR model. These models had similar performance when predicting the ages of groups 14, 21, 28, and 35 days.

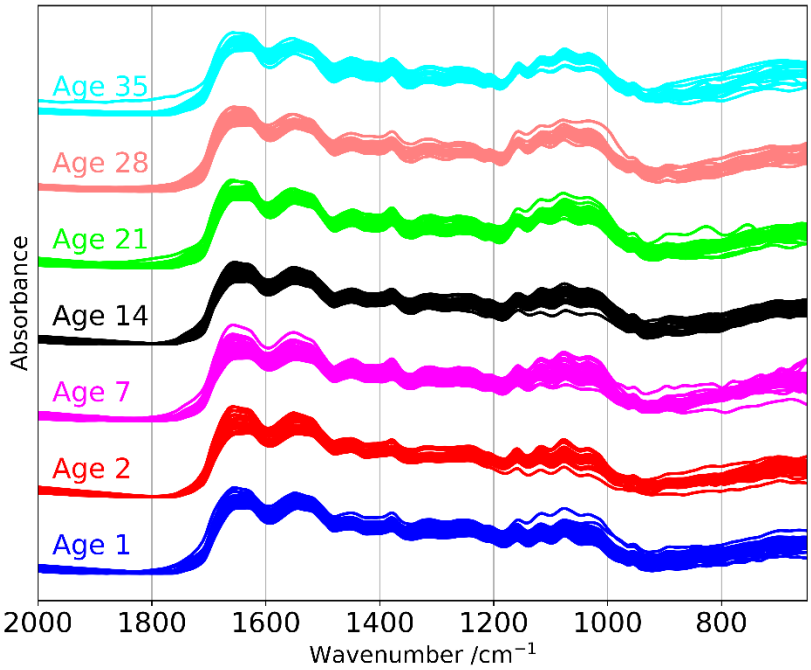

Figure 13. Overlay of raw female *Ae. triseriatus* spectra separated by age where ages are distinguished by color as indicated in the legend and offset by an integer constant

When validating the performance of the PLSR and ANN models, autocorrelation was not taken into account. In a time series, autocorrelation refers to earlier observations having effects on later observations in the time series (e.g., removing mosquitoes at day 7 adding stress to

mosquitoes left in the colony which is measurable in later groups in the time series). The mosquitoes were treated as separate entities within the colony such that one mosquito had no effect on the other mosquitoes. Assuming no autocorrelation is consistent with prior studies using PLSR for mosquito age prediction.[55] Therefore, our calculations for SEPsv are likely optimistic compared to the true predictive performance of these models. Additional validation should be conducted to investigate the impact of autocorrelation on the predictive model.

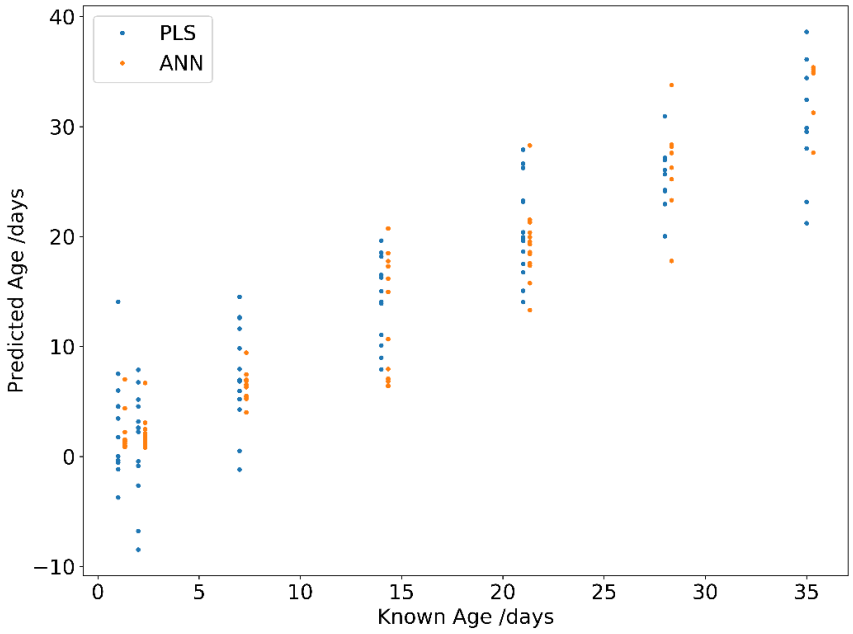

Figure 14. Model performance of PLS and artificial neural networks (ANN) in a scatter plot with PLS represented by blue points and ANN represented by orange points. Models are staggered by 1/3 day on the x-axis.

A line of best fit with 95% confidence intervals for predicted age by the PLSR model versus known age is shown in Figure 15. Meaning that there is a 95% probability that the true linear regression line modeling predicted age versus known age for the PLSR model is within the confidence interval represented by lines above and below the line of best fit. This line of best fit has a slope of 0.8721 and y-intercept of 1.068. The coefficient of determination or R^2 value is

0.8682 meaning that the line of best fit explains 86.82% of the variation in age predicted by the PLSR model. An R^2 value of 1 would mean that all of the variance in predicted age is explained by the line of best fit whereas an R^2 value of 0 would mean the none of the variance in predicted age is explained by the line of best fit. A line of best fit with 95% confidence intervals for predicted age by the ANN model versus known age is shown in Figure 16. The slope of the line of best fit is 0.9412 with a y-intercept of 0.1352 and R^2 value of 0.9231. Looking at the slopes of the regression line, or line of best fit, for the PLSR and ANN models, a slope closer to 1 indicates a higher performance for the ANN model because predicted age and known age should increase at the same rate. Additionally, a y-intercept closer to 0 indicates better performance for the ANN model because if the known age is 0 days, there should be no age to predict.

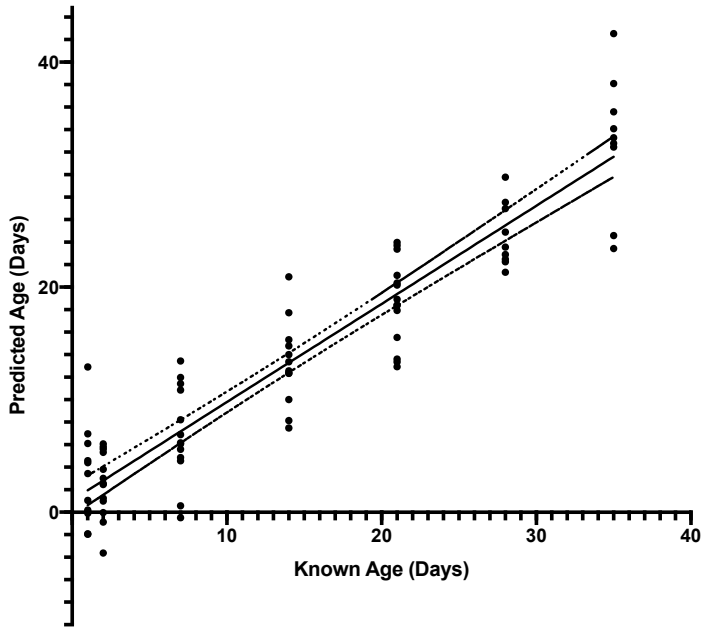

Figure 15. Line of best fit with 95% confidence intervals for age predicted by the PLSR model versus known age. The slope of the line of best fit is 0.8721 with a y-intercept of 1.068 and an R^2 value of 0.8682.

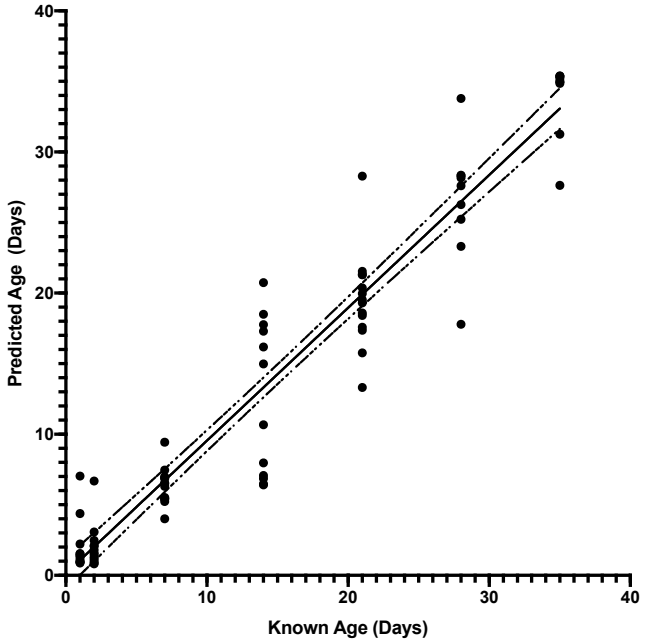

Figure 16. Line of best fit with 95% confidence intervals for age predicted by the PLSR model versus known age. The slope of the line of best fit is 0.9412 with a y-intercept of 0.1352 and an R^2 value of 0.9231.

To better understand the chemical differences of these mosquitoes, the mean normalized spectra for each age group was plotted and shown in Figure 17. The greatest difference in absorbance is between 1150 cm^{-1} and 1000 cm^{-1}. The series of peaks in this region are typically associated with chitin—a polymer comprised of polysaccharides which serves as a primary component in the exoskeleton of mosquitoes or other arthropods shown in Figure 18.[56] The chitin bands are prominent in the infrared spectra and therefore play an important role when being fit by the models. A zoomed view of this region is shown in Figure 19 and the band at 1032 cm^{-1} is indicated by the vertical line. The band absorbances increase as a function of age with the exception of ages 1 and 2 days which are out of order. This trend can be more easily observed in a plot of band height at 1032 cm^{-1} versus age in Figure 20. Since data were pre-processed with a second derivative function, the second derivative of absorbance at 1032 cm^{-1}

was plotted versus known age in days shown in Figure 21 with a more clear linear relationship.

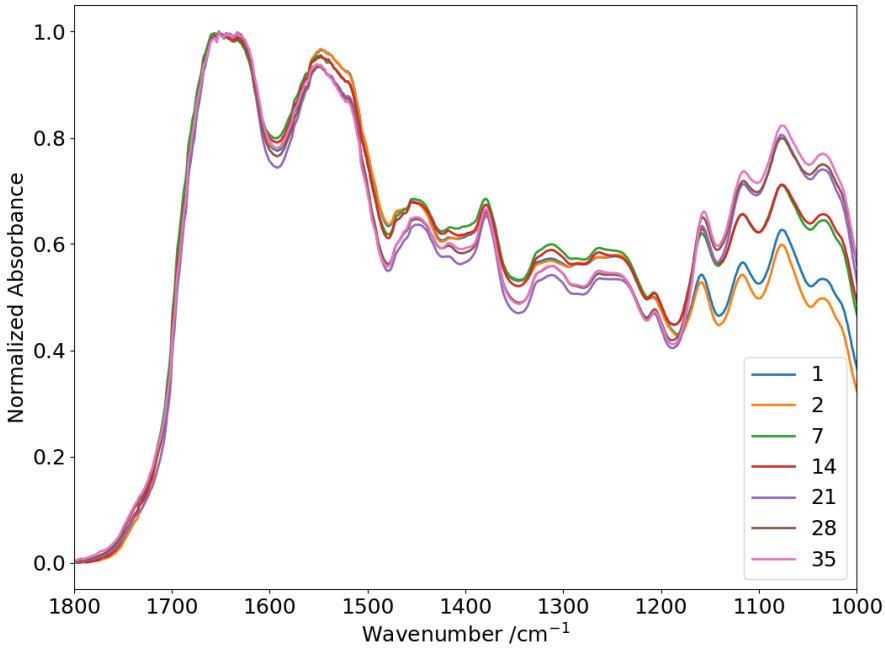

Figure 17. Plot of normalized mean spectra for each age group cropped to the region of 1800 – 1000 cm^{-1} where normalized absorbance is on the y-axis and wavenumber (cm^{-1}) is on the x-axis.

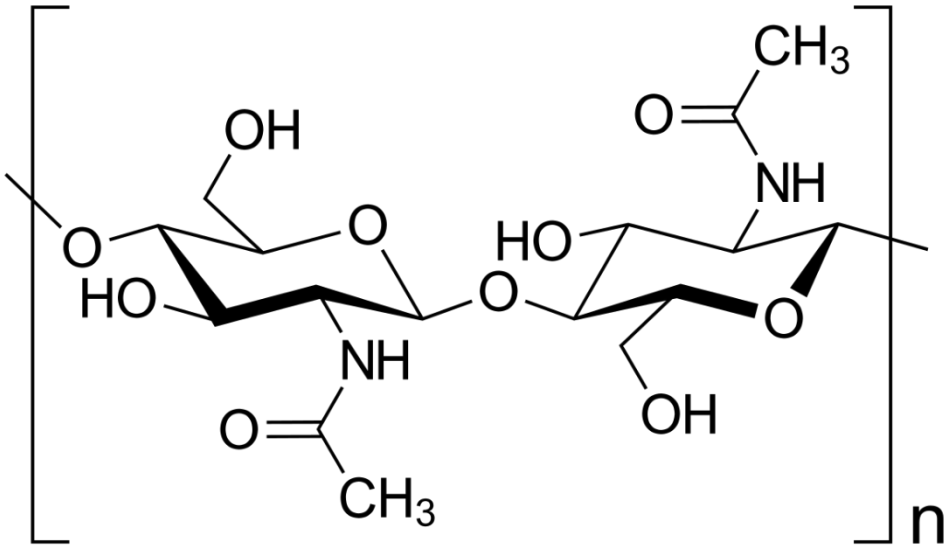

Figure 18. Structure of monomeric unit that makes up the polysaccharide chitin.

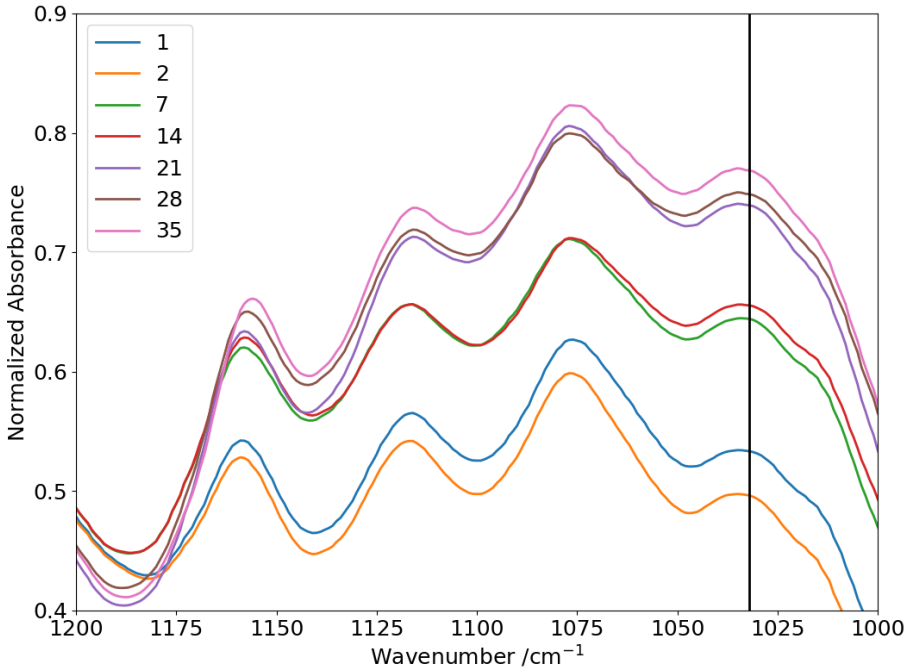

Figure 19. Plot of normalized mean spectra for each age group cropped to 1200 – 1000 cm^{-1} of normalized absorbance versus wavenumber. The vertical black line is at 1032 cm^{-1} where the peak height generally increases as the age group increases, except for age groups 1 & 2.

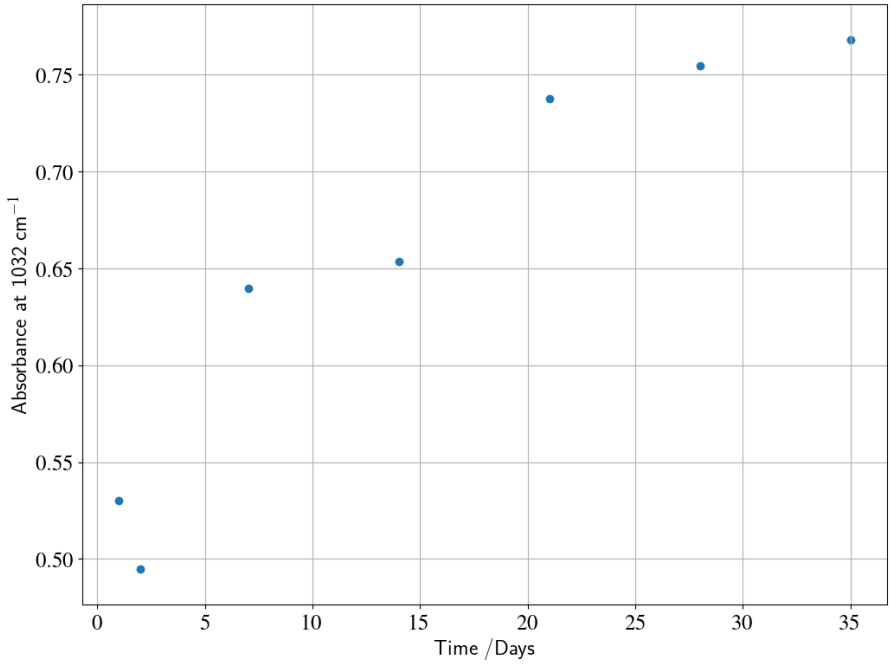

Figure 20. Scatter plot of mean absorbance at 1032 cm^{-1} for each age group versus time in days.

The ANN model had a lower SEPsv than the PLSR model. PLSR works best with linear or near linear relationships, so it makes sense that it would have trouble modeling this trend we see with the peak height at 1032 cm^{-1}. From Figure 14 and Figure 16 it is evident that the PLSR model does much better at predicting age after the first two age groups. ANN is a much more sophisticated model and excels at modelling non-linear complex data. From Figures 15 and 16, it appears as though the ANN model was able to overcome the trend reversal in peak height at 1032 cm^{-1} as the predictions for age groups 1 and 2 are much better when compared with the age predictions made by PLSR. The ANN model's performance decreases after the first three groups and performs similar to the PLSR model.

The observation of the trend at band 1032 cm^{-1} provides information about the biochemistry in the mosquitoes as they age. In order to metamorphose into adult mosquitoes from pupae, the pupae have to molt.[57] Once mosquito pupae shed their skin and metamorphose into adults, they are very fragile and when handling. It was very easy to damage their limbs when transferring mosquitoes to a separate container to kill via freezing or when removing their legs for IR measurement. After a few days, mosquitoes have seemingly strengthened their exoskeleton and are much more resistant to limb damage. Since there is a linear relationship between concentration of a chemical and absorbance in spectroscopy via Beer's Law, more chitin in the exoskeleton would result in higher absorbance.[58] The nonlinearity of this trend in days 1 and 2 could be explained by excess chitin expressed during the pupal phase of development. Generally, chitin expression increases through different life stages in order to build up a protective exoskeleton. There may be a change in where chitin is being expressed in different parts of the mosquito as it tries to develop its exoskeleton after molting.[57]

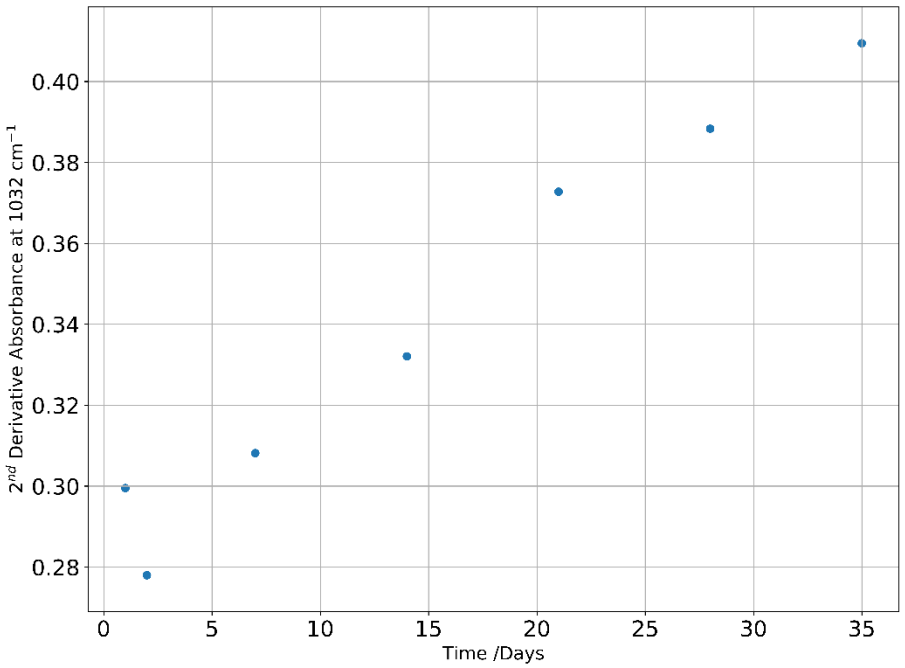

Figure 21. Scatter plot of second derivative of average absorbance at 1032 cm^{-1} for each age group versus time in days (*Aedes triseriatus*).

It may be possible to use chitin as a biomarker for age prediction. The expression of chitin as a function of age in adult female mosquitoes has been investigated using qPCR.[57] It may be possible to develop a method using infrared spectroscopy that focuses only on chitin absorbance bands. For this, a surface technique such as ATR might be more appropriate. Furthermore, it has been shown that cuticular protein expression can be used to determine the age of mosquitoes.[19]

The PLSR and ANN predictive models were used to try and predict the ages of *Culex tarsalis* mosquitoes used in Study 2 without making any modifications to the methods. Neither model was able to predict the age of *Cx. tarsalis* mosquitoes. The second derivative of mean absorbance for each age bin at 1032 cm^{-1} was plotted versus known age in days shown in Figure 22 and shows a linear relationship between the second derivative of absorbance and known age in days. Results from this investigation of *Culex tarsalis* further suggest the possibility of

modeling mosquito age while focusing on the region between 1200 and 1000 cm^{-1} where chitin is prominent in the mid-infrared region of the electromagnetic spectrum.

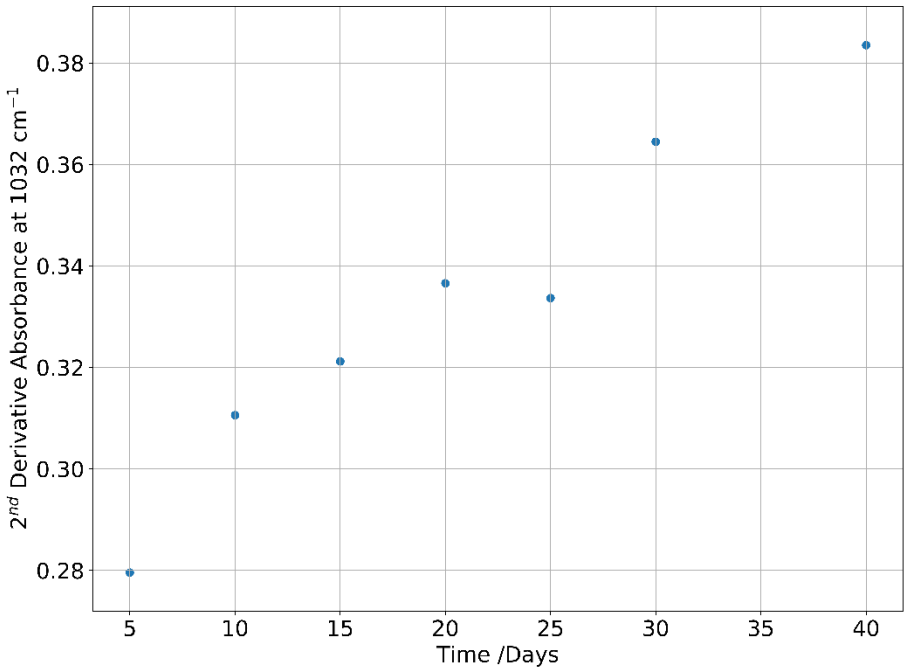

Figure 22. Scatter plot of second derivative of average absorbance at 1032 cm^{-1} for each age group versus time in days (*Culex tarsalis*).

CHAPTER FOUR: CONCLUSIONS & FUTURE DIRECTIONS

Using mid-infrared spectroscopy and chemometrics, the chronologic age of *Ae. triseriatus* mosquitoes were predicted using a PLSR and ANN model. *Aedes triseriatus* were successfully reared into groups of different ages with low uncertainty in the age. *Aedes triseriatus* spectra were used create a training dataset to fit models for prediction using PLSR and ANN. PLSR and ANN models were used to predict the age of samples using a validation dataset with SEPsv of 4.3 and 3.3 days respectively. Mean spectra for each age group were used to try and discern a specific chemical underpinning for the performance of these models and to explain why mosquito age could be predicted using PLSR and ANN models. Peaks between 1200 – 1000 cm^{-1} typically associated with chitin were investigated and the second derivative of mean absorbance by age at 1032 cm^{-1} increased linearly with age for both *Aedes triseriatus* and *Culex tarsalis*.

Future experiments should be conducted to (1) further validate the performance of these PLSR and ANN models used for mosquito age prediction, (2) better understand how mosquitoes' biochemistries change as a function of age, and (3) investigate the correlation between chitin and mosquito age. The *Ae. triseriatus* time series study should be repeated and the PLSR and ANN models should attempt to predict the ages of these mosquitoes. Moreover, the PLSR and ANN models can be improved by expanding the size of the training datasets to encompass greater variation of samples at each age. These models could also be improved by conducting additional time course studies with different age bins to make the known ages data more continuous and less categorical. Similar time series experiments should be conducted using a different species of mosquito—ideally one within the same genus, *Aedes*, and one with a different genus, such as *Culex*, in order to test the universality of using IR spectroscopy and chemometrics. Depending

on the performance of the PLSR and ANN models when predicting the ages of species other than *Ae. triseriatus*, new models should be developed for the different species or new training datasets should be developed that includes multiple species. Lastly, investigating the correlation between chitin and mosquito age should be conducted using an infrared surface technique such as ATR. A model that focuses on changes in chitin as a function of age may serve as an alternative method for predicting the chronologic age of mosquitoes.

REFERENCES

[1] WHO. A global brief on vector-borne diseases. Technical report, World Health Organization, 2014.

[2] AMCA. Best Practices for Mosquito Control: a Focused Update. Technical report, American Mosquito Control Association. 2017.

[3] Classification of Mosquitoes with Infrared Spectroscopy and Partial Least Squares-Discriminant Analysis, Lamye Sroute, Brian D. Byrd, and Scott W. Huffman.

[4] Markowski, D. (2015, August 5) The Key Components of an Integrated Mosquito Management Program. *Vector Disease Control International.*

[5] Davies, M.; Foust, E.; Williams, C.; Watkins, H.; Michael, L.; Zimmerman, S. North Carolina Vector Borne Disease Management. 2016; http://epi.publichealth.nc.gov/cd/vector/VectorborneDiseaseProgramWhitePaper.pdf.

[6] Eldridge, B. F. Strategies for surveillance, prevention, and control of arbovirus diseases in western North America. American Journal of Tropical Medicine and Hygiene 1987, 37, 77–86.

[7] Grard, G.; Caron, M.; Mombo, I. M.; Nkoghe, D.; Mboui Ondo, S.; Jiolle, D.; Fonte-nille, D.; Paupy, C.; Leroy, E. M. Zika Virus in Gabon (Central Africa) 2007: A New Threat from Aedes albopictus? PLOS Neglected Tropical Diseases 2014, 8, 1–6.

[8] Li, C. X.; Guo, X. X.; Deng, Y. Q.; Xing, D.; Sun, A. J.; Liu, Q. M.; Wu, Q.; Dong, Y. D.; Zhang, Y. M.; Zhang, H. D.; Cao, W. C.; Qin, C. F.; Zhao, T. Y. Vector competence and transovarial transmission of two Aedes aegypti strains to Zika virus. Emerging Microbes and Infections 2017, 6, 1–7.

[9] Bhatt, S.; Gething, P.; Brady, O.; Messina, J.; Farlow, A.; Moyes, C. The global distribution

and burden of dengue. Nature 2012, 496, 504–507.

[10] Hernandez-Triana, L. M.; Jeffries, C. L.; Mansfield, K. L.; Carnell, G.; Fooks, A. R.; Johnson, N. Emergence of West Nile Virus Lineage 2 in Europe: A Review on the Introduction and Spread of a Mosquito-Borne Disease. Frontiers in Public Health 2014, 2, 1–8.

[11] Bewick, S.; Agusto, F.; Calabrese, J. M.; Muturi, E. J.; Fagan, W. F. Epidemiology of LaCrosse Virus Emergence, Appalachia Region, United States. Emerging Infectious Diseases 2016, 22, 1921–1929.

[12] Thiboutot, M. M.; Kannan, S.; Kawalekar, O. U.; Shedlock, D. J.; Khan, A. S.; Sarangan, G.; Srikanth, P.; Weiner, D. B.; Muthumani, K. Chikungunya: A potentially emerging epidemic? PLoS Neglected Tropical Diseases 2010, 4, 1–8.

[13] NACCHO Report: Mosquito Control Capabilities in the U.S., 2017.

[14] Zhu, F.; Lavine, L.; O'Neal, S.; Lavine, M.; Foss, C.; Walsh, D. Insecticide Resistance and Management Strategies in Urban Ecosystems. Insects 2016, 7 (1), 2 DOI: 10.3390/insects7010002.

[15] Tjaden, N. B.; Thomas, S. M.; Fischer, D.; Beierkuhnlein, C. Extrinsic Incubation Period of Den-gue: Knowledge, Backlog, and Applications of Temperature Dependence. PLoS Neglected Trop-ical Diseases 2013, 7 (6) DOI: 10.1371/journal.pntd.0002207.

[16] Mosquito-borne Transmission https://www.cdc.gov/dengue/training/cme/ccm/page45915.html.

[17] Hugo, L. E., Quick-Miles, S., Kay, B. H., and Ryan, P. A. (2008) Evaluations of Mosquito Age Grading Techniques Based on Morphological Changes. *Journal of Medical Entomology 45*, 353–369.

[18] Müller, P.; Pflüger, V.; Wittwer, M.; Ziegler, D.; Chandre, F.; Simard, F.; Lengeler, C. Identification of Cryptic Anopheles Mosquito Species by Molecular Protein Profiling. PLoS ONE 2013, 8.

[19] Cook, P. E., and Sinkins, S. P. (2010) Transcriptional profiling of Anopheles gambiae mosquitoes for adult age estimation. *Insect Molecular Biology 19*, 745–751.

[20] Kesavaraju, B.; Lampman, R. L.; Krasavin, N. M.; Hutchinson, M.; Graves, S. E.; Dickson, S. L.; Farajollahi, A. Evaluation of a Rapid Analyte Measurement Platform for West Nile Virus Detection Based on United States Mosquito Control Programs. The American Journal of Tropical Medicine and Hygiene 2012, 87 (2), 359–363 DOI: 10.4269/ajtmh.2012.11-0662.

[21] Agelet, L. E.; Hurburgh, C. R. A tutorial on near infrared spectroscopy and its calibration. Critical Reviews in Analytical Chemistry 2010, 40, 246–260.

[22] Moudgil, H. K. Textbook of Physical Chemistry; PHI Learning Private Limited: Delhi, 2014.

[23] Krajacich, B. J.; Meyers, J. I.; Alout, H.; Dabire, R. K.; Dowell, F. E.; Foy, B. D. Analysis of near infrared spectra for age-grading of wild populations of Anopheles gambiae. Parasites & Vectors 2017, 10, 552–565.

[24] Sikulu-Lord, M. T., Milali, M. P., Henry, M., Wirtz, R. A., Hugo, L. E., Dowell, F. E., and Devine, G. J. (2016) Near-Infrared Spectroscopy, a Rapid Method for Predicting the Age of Male and Female Wild-Type and Wolbachia Infected Aedes aegypti. *PLOS Neglected Tropical Diseases 10.*

[25] Reeves, L. E.; Holderman, C. J.; Blosser, E. M.; Gillett-Kaufman, J. L.; Kawahara, A. Y.; Kaufman, P. E.; Burkett-Cadena, N. D. Identification of Uranotaenia sapphirina as a

specialist of annelids broadens known mosquito host use patterns. Communications Biology 2018, 1 (1) DOI: 10.1038/s42003-018-0096-5.

[26] Childs, L. M.; Cai, F. Y.; Kakani, E. G.; Mitchell, S. N.; Paton, D.; Gabrieli, P.; Buckee, C. O.; Catteruccia, F. Disrupting Mosquito Reproduction and Parasite Development for Malaria Control. PLOS Pathogens 2016, 12 (12) DOI: 10.1371/journal.ppat.1006060.

[27] Harbach, R. Family Culicidae Meigen, 1818 http://mosquito-taxonomic-inventory.info/family-culicidae-meigen-1818.

[28] Brady, O. J.; Godfray, H. C. J.; Tatem, A. J.; Gething, P. W.; Cohen, J. M.; Mckenzie, F. E.; Perkins, T. A.; Reiner, R. C.; Tusting, L. S.; Sinka, M. E.; et al. Vectorial capacity and vector control: reconsidering sensitivity to parameters for malaria elimination. Transactions of The Royal Society of Tropical Medicine and Hygiene 2016, 110 (2), 107–117 DOI: 10.1093/trstmh/trv113.

[29] Barker, C. M.; Eldridge, B. F.; Reisen, W. K. Seasonal Abundance of Culex tarsalis and Culex pipiens Complex Mosquitoes (Diptera: Culicidae) in California. Journal of Medical Entomology 2010, 47 (5), 759–768 DOI: 10.1603/me09139.

[30] Szumlas, D. E.; Apperson, C. S.; Powell, E. E.; Hartig, P.; Francy, D. B.; Karabotsos, N. Relative Abundance and Species Composition of Mosquito Populations (Diptera: Culicidae) in a La Crosse Virus-Endemic Area in Western North Carolina. Journal of Medical Entomology 1996, 33 (4), 598–607 DOI: 10.1093/jmedent/33.4.598.

[31] Baker, M. J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols 2014, 9, 1771–1791.

[32] Wilson, E. B.; Decius, J. C.; Cross, P. C. Molecular vibrations: the theory of infrared and raman vibrational spectra; Dover Publications: New York, 1955.

[33] Stuart, B. H. Infrared Spectroscopy: Fundamentals and Applications; John Wiley & Sons, 2005; Vol. 1; pp 25–33, 35.

[34] Kazarian, S.; Chan, K. Applications of ATR-FTIR spectroscopic imaging to biomedical samples. Biochimica et Biophysica Acta – Biomembranes 2006, 1758, 858 – 867

[35] Mitchell, M. B. Structure-Property Relations in Polymers; American Chemical Society, 1993; Vol. 236; pp 351–375.

[36] Khoshmanesh, A.; Christensen, D.; Perez-Guaita, D.; Iturbe-Ormaetxe, I.; O'Neill, S. L.; McNaughton, D.; Wood, B. R. Screening of Wolbachia endosymbiont infection in Aedes aegypti mosquitoes using attenuated total reflection mid-Infrared spectroscopy. Analytical Chemistry 2017, 89, 5285–5293.

[37] Theophilou, G.; Lima, K. M. G.; Martin-Hirsch, P. L.; String fellow, H. F.; Martin, F. L. ATR-FTIR spectroscopy coupled with chemometric analysis discriminates normal, borderline and malignant ovarian tissue: classifying subtypes of human cancer. The Analyst 2016, 141, 585–594.

[38] Davis, R.; Mauer, L. Fourier transform infrared (FT-IR) spectroscopy: A rapid tool for detection and analysis of foodborne pathogenic bacteria. Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology. 2010, 2, 1582–1594.

[39] Zhang, Y. P.; Lewis, R.; McElhaney, R. N.; Hodges, R. S. Interaction of a peptide model of a hydrophobic transmembrane α-helical segment of a membrane protein with phosphatidyl-choline bilayers: Differential scanning calorimetric and FTIR spectroscopic studies. Biochemistry 1992, 31, 11579–11588.

[40] Cozzolino, D.; Holdstock, M.; Dambergs, R. G.; Cynkar, W. U.; Smith, P. A. Mid infrared

spectroscopy and multivariate analysis: A tool to discriminate between organic and non-organic wines grown in Australia. Food Chemistry 2009, 116, 761–765.

[41] Al-Jowder, O.; Defernez, M.; Kemsley, E. K.; Wilson, R. H. Mid-Infrared Spectroscopy and Chemometrics for the Authentication of Meat Products. Journal of Agricultural and Food Chemistry 1999, 47 (8), 3210–3218 DOI: 10.1021/jf981196d.

[42] Zarnowiec, P.; Lechowicz, L.; Czerwonka, G.; Kaca, W. Fourier transform infrared spectroscopy (FTIR) as a tool for the identification and differentiation of pathogenic bacteria. Current Medicinal Chemistry 2015, 22, 1710–1718.

[43] Movasaghi, Z.; Rehman, S.; Rehman, I. U. Fourier transform infrared (FTIR) spectroscopy of biological tissues. Applied Spectroscopy Reviews 2008, 43, 134–179.

[44] Vivó-Truyols, G.; Schoenmakers, P. J. Automatic selection of optimal Savitzky-Golay smoothing. Analytical Chemistry 2006, 78, 4598–4608.

[45] Cao, J.; Ng, E. S.; McNaughton, D.; Stanley, E. G.; Elefanty, A. G.; Tobin, M. J.; Heraud, P. Using Fourier transform IR spectroscopy to analyze biological materials. Journal of Biophotonics 2014, 7, 767–781.

[46] Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 1964, 36 (8), 1627–1639 DOI: 10.1021/ac60214a047.

[47] Haaland, D. M.; Thomas, E. V. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical Chemistry 1988, 60, 1193–1202.

[48] Shaw P.J.A. (2003) *Multivariate statistics for the Environmental Sciences*, Hodder-Arnold. ISBN 0-340-80763-6.

[49] Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001, 58 (2), 109–130 DOI: 10.1016/s0169-7439(01)00155-1.

[50] Barker, M.; Rayens, W. Partial least squares for discrimination. Journal of Chemometrics 2003, 17 (3), 166–173 DOI: 10.1002/cem.785.

[51] Marini, F.; Bucci, R.; Magrì, A.; Magrì, A. Artificial neural networks in chemometrics: History, exam-ples and perspectives. Microchemical Journal 2008, 88 (2), 178–185 DOI: 10.1016/j.microc.2007.11.008.

[52] Kauffman, E., Payne, A., Franke, M. A., Schmid, M. A., Harris, E. and Kramer, L. D. (2017). Rearing of *Culex* spp. and *Aedes* spp. Mosquitoes. *Bio-protocol* 7(17): e2542. DOI: 10.21769/BioProtoc.2542.

[53] Weston, S.; Frieman, M. B. COVID-19: Knowns, Unknowns, and Questions. mSphere 2020, 5 (2) DOI: 10.1128/msphere.00203-20.

[54] The HDF5® Library & File Format https://www.hdfgroup.org/solutions/hdf5/.

[55] Sikulu, M.; Killeen, G. F.; Hugo, L. E.; Ryan, P. A.; Dowell, K. M.; Wirtz, R. A.; Moore, S. J.; Dowell, F. E. Near-infrared spectroscopy as a complementary age grading and species identification tool for African malaria vectors. Parasites & Vectors 2010, 3 (1) DOI: 10.1186/1756-3305-3-49.

[56] Kumirska, J.; Czerwicka, M.; Kaczyński, Z.; Bychowska, A.; Brzozowski, K.; Thöming, J.; Stepnowski, P. Application of Spectroscopic Methods for Structural Analysis of Chitin and Chitosan. Marine Drugs 2010, 8 (5), 1567–1636 DOI: 10.3390/md8051567.

[57] Yang, X.; Yin, Q.; Xu, Y.; Li, X.; Sun, Y.; Ma, L.; Zhou, D.; Shen, B. Molecular and physiological characterization of the chitin synthase B gene isolated from Culex pipiens

pallens (Diptera: Culicidae). Parasites & Vectors 2019, 12 (1) DOI: 10.1186/s13071-019-3867-z.

[58] Beer. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. Annalen der Physik und Chemie [Determination of the absorption of red light in colored liquids] 1852, 162 (5), 78–88 DOI: 10.1002/andp.18521620505.

APPENDIX

Glossary of Biologically Relevant Terms

arbovirus – viruses transmitted by things like mosquitoes, ticks, or arthropods

eclosion – when pupae metamorphose or transition to adult mosquitoes

epidemiology – branch of medicine related to disease control

extrinsic incubation period – time between acquisition of a pathogen and when the pathogen can be passed onto a susceptible host

instar stages – different development stages of mosquito larvae

larvae – first stage of development for mosquitoes after hatching, small and aquatic

nulliparous – hasn't laid eggs yet

oogenesis – the development of egg cells into competent cells capable of further development when fertilized

oviposit – to lay (eggs)

parous – has laid eggs

pupae – second stage of development after larvae, still aquatic, but bigger